Future of Life Institute
PO Box 706
Allston, MA, 02134

Additional Comments on the "White Paper: On Artificial Intelligence - A European approach to excellence and trust"

I. Introduction

The Future of Life Institute (FLI) is a U.S. based, but globally focused, non-profit working at the intersection of emerging technologies like artificial intelligence (AI) and governance. For example, FLI helped organize the creation of one of the earliest and most influential set of principles on the development and governance of AI, the Asilomar AI Principles.[1] FLI has also supported AI safety research and organized annual conferences bringing together hundreds of the world's top AI researchers to address key challenges in responsible development trajectories. More recently, FLI was honored to participate as a "champion" for the U.N. Secretary-General's Digital Cooperation Roadmap, advising on AI-related global governance issues, along with two European Union Member States and the Governments of France and Finland.[2] With this perspective, we commend the European Commission for pursuing a positive, proactive governance approach to ensuring the trustworthy development and deployment of AI in Europe.

We also recognize, however, that the Commission may receive many concerns and comments from companies that produce AI systems for the European market (or their association representatives). We encourage the Commission to give thoughtful consideration to their input, but urge the Commission *not* to weaken its regulatory approach to AI, as many might suggest should happen. We also strongly believe that a successful European approach to trustworthy AI depends on it being forward-looking and prospectively adaptable to governance challenges presented by future technical improvements to AI systems. This view is endorsed by many leading AI researchers and AI policy experts across the globe in an open letter, available here.[3]

---

[1] Created at an FLI organized workshop in 2017, the Asilomar AI Principles are signed by over 1,500 leading AI and robotics researchers, and over 3,500 other prominent individuals. For more, see: https://futureoflife.org/ai-principles/.
[2] See, United Nations, "Report of the Secretary-General: Roadmap for Digital Cooperation" 06.2020, available at https://www.un.org/en/content/digital-cooperation-roadmap/.
[3] See the open letter, "Apply Foresight in a Meaningful Regulatory Approach to AI," available at http://futureoflife.org/foresight-in-ai-regulation-open-letter.

Though we are broadly very supportive of the White Paper, we offer the below (summarized) list of recommendations to the Commission to further strengthen the stated goals. These recommendations are explained in the remaining sections of our written submission. We would happily answer any further questions the Commission may have about these recommendations, and look forward to future opportunities to support European governance of AI.

We recommend that the Commission:

- Require periodic safety reassessments of continual self-learning AI systems;
- Resolve legal uncertainty related to "stand-alone" software;
- Restrict use of the development risk defence and later defect defence for liability protection on continual self-learning AI systems;
- Create an obligation to monitor for continual self-learning AI systems;
- Include consideration of societal harms in risk assessments
- Create a multi-tiered risk assessment framework;
- Evaluate supplemental governance methods to accompany a voluntary labelling scheme;
- Establish reciprocal legal responsibilities for AI systems;
- Mandate disclosures of conflicts of interests in AI systems; and
- Evaluate AI "loyalty" as an immaterial risk in conformity assessments.

II. Key Recommendations

A. Recommendations for Safety Reviews and Liability for Continual Self-Learning AI Systems During the Product Life Cycle

We agree with the Commission that "the use of AI in products and services can give rise to risks that EU legislation currently does not explicitly address…" and that those risks "... may be present at the time of placing products on the market or arise as a result of software updates or is self-learning when the product is being used."[4] If properly implemented, the risk management processes identified in the White Paper may mitigate these risks for AI products and services before being placed on the market. See our recommendations for improving the risk management process in Section II.B of this document. However, as also correctly identified by the Commission, we believe that there "may be also situations in the future where the outcomes of the AI systems cannot be fully determined in advance. In such a situation, the risk assessment performed before placing the product on the market may no longer reflect the use, functioning or behaviour of the product."[5]

[4] European Commission, "White Paper: On Artificial Intelligence - A European approach to excellence and trust," EN, 19.2.2020, p. 14.
[5] European Commission, "Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics," EN, 19.2.2020, p. 7.

This unique regulatory challenge presented by AI systems is particularly problematic for self-learning systems that continue to learn throughout their product life cycle. By "product life cycle" we refer to the stage of a product after it has been placed on the market and is only subject to market surveillance for enforcement. This meaning derives from the Commission's provided diagram on product safety legislation.[6] For simplicity, we will refer to these as ***continual self-learning AI systems***.[7] Continual self-learning AI systems often use a technique referred to as online machine learning, which allows a product or service to adapt and evolve to a user's preferences or continue to self-improve through the data gathered during its use. There is a further distinction between continual self-learning AI systems that must be made for regulatory purposes, based on the degree to which the AI system operates from a *centralized* or *decentralized* AI model. In a ***centralized*** continual self-learning AI product or service, there is, in essence, a single AI model that is self-learning, and that self-learning on a single model is applied through a network to all instances in which it is used in a product. For example, consider a voice assistant device produced by Company X. Each voice assistant device has an integrated continual self-learning AI system designed to recognize speech and respond to queries through its speaker and microphone. The AI system is continuing to self-learn after product introduction whenever any of its prospective customers uses it, but is doing so through a networked, centralized model which its producer, Company X, can maintain insight into and potential oversight of as it self-learns. In a ***decentralized*** continual self-learning AI product or service, the system continues to learn, but learns uniquely in each or some of its instances of deployment. For example, consider a hypothetical AI-enabled cleaning robot. Though each cleaning robot produced by Company Y is trained on a similar data set, once purchased by a customer, the cleaning robot's AI model becomes decentralized from every other cleaning robot, perhaps because it is intentionally disconnected from a cloud network for privacy or cybersecurity considerations. However, the cleaning robot continues to self-learn based on the customer's preferences and characteristics of its deployed office environment (e.g., the layout of the physical space unique to each office, how long to let the half-filled coffee cup stay on the counter before cleaning). Company Y has less transparency into how each cleaning robot

[6] See the Commission's diagram on "The underlying logic of the current Union product safety legislation" found in the "Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics," p. 5.

[7] FLI has chosen the term "continual self-learning AI systems" to best convey meaning to a policy-orientated audience. However, In the technical machine learning literature, "online learning" generally refers to algorithms and mechanisms that learn sequentially and incrementally rather than in a batch of data. This technique is most commonly used after deployment in live environments. Used in this context, FLI does **not** intend to invoke the distinct, technical variation of "continual learning." This term generally references a more specific kind of online learning that meets additional criteria, such as the expectation that the system is learning new skills in a post-deployment context. For our non-technical audience, we choose the term "continual self-learning AI" to mean those systems that are learning either on an ongoing or periodic batched basis in the post-deployment environment, as it is a convenient term for the present audience. For a description of "continual learning" in a more technical context, see, for example:
Parsi, G., Kemker, R., Part, J.L., Kanan, C., Wermter, S., "Continual lifelong learning with neural networks: A review," 05.2019, Neural Networks, available at https://doi.org/10.1016/j.neunet.2019.01.012

continues to learn unless allowed by the customer, and the AI models might start to diverge considerably across all instances of the product based on their self-learning.

FLI believes the Commission must develop specific pieces of legislation that address how the safety of centralized and decentralized continual self-learning products and services are reevaluated, and how liability is maintained for these systems. To be clear, continual self-learning products and services are not particularly common, *today*, in the European or other markets. However, the Commission is absolutely correct that "given how fast AI is evolving, the regulatory framework must leave room to cater for further developments."[8] Continual self-learning AI products and services is one such further development that will likely be used with increasing frequency in the future. Therefore, it would be negligent and short-sighted for the Commission to embark on any revisions to a legal framework that do *not* address these and other types of AI systems likely to be more prevalent in the future. **This view is supported by the many AI researchers and policy experts from across the globe who have signed a letter supporting the Commission taking a meaningful, future-oriented approach regarding the effects of AI systems on the rights and safety of EU citizens.[9]**

To that end, FLI has several recommendations for improving upon the already robust approach the Commission has set forth.

Recommendation: Require periodic safety reassessments of continual self-learning AI systems

FLI strongly agrees with the Commission that "particular account should be taken of the possibility that certain AI systems evolve and learn from experience, which may require repeated assessments over the life-time of the AI systems in question."[10] Therefore, we recommend the Commission require continual self-learning AI products and services resubmit for conformity assessment after a period of time determined by officials during the initial, or previous, conformity assessment. Allowing regulatory authorities to determine how frequently to reassess the continued safety of continual self-learning AI will provide the necessary discretion to regulatory officials. These officials will need to respond to a multitude of highly specific variables for each continual self-learning AI system that may make a system more or less likely to evolve in risky ways, thus requiring different frequency of reassessment. These variables include, but are not limited to, the amount of new data and continued learning that is occurring by particular AI systems in their use environment, the degree to which this new learning environment comports to the trained learning environment, whether there is reason to expect the system will be intentionally manipulated through adversarial examples or subject to

[8] European Commission, "White Paper: On Artificial Intelligence - A European approach to excellence and trust," EN, 19.2.2020, p. 10.
[9] Please see the open letter, "Apply Foresight in a Meaningful Regulatory Approach to AI," available at http://futureoflife.org/foresight-in-ai-regulation-open-letter.
[10] European Commission, "White Paper: On Artificial Intelligence - A European approach to excellence and trust," EN, 19.2.2020, p. 23.

malicious use, whether recent AI safety research has discovered particular new concerns for using prior "state of the art" techniques, and so forth. These conditions and others will lead to variable amounts of safety risk for a particular continual self-learning AI product and services that can be determined in the conformity assessment, therefore requiring variable frequency in reassessments.

Notably, it would be uniquely more difficult for a decentralized continual self-learning product or service to be resubmitted for a conformity assessment, as there is no longer a single model to reassess, but many various iterations that evolved from an original model. Each iteration of the decentralized continual self-learning AI system is held privately by different customers of the original producer, and may contain sensitive data or information. Solutions to this problem have not been robustly evaluated, though FLI believes strongly that the Commission could develop one. For example, as a condition of sale, such as through a warranty or legal contract, it could be possible for the producer to require that a random sample of such instances would need to be anonymously audited in the future for continued safety. Alternatively, for high-risk applications, the legal responsibility to maintain the safety of the decentralized continual self-learning AI systems may transfer between the producer and customer in a clearly understood and transparent manner. It would then become the customer's responsibility to have the system evaluated for continued conformity with EU legal frameworks. We acknowledge that further research would need to be conducted to design a regulatory approach that sufficiently addresses the needs of producers and customers of decentralized continual self-learning AI systems under such proposals to not place unnecessary regulatory burden on either. However, we are confident such a regulatory approach, or a set of alternative approaches that address the identified fundamental problem,[11] can be developed to ensure the safety of AI in Europe.

Recommendation: Resolve legal uncertainty related to "stand-alone" software

The European Commission has properly identified that considerable uncertainty remains whether stand-alone software (which could include stand-alone continual self-learning AI systems) are "products" or "services" and thus included in the relevant safety and liability directives.[12] Likewise, this presents any number of obvious problems the Commission has already identified, including that uncertainty can "reduce overall levels of safety and undermine

[11] For example, in a later recommendation to "Create a Obligation to Monitor continual Self-learning AI Systems," FLI proposes one such alternative that would mitigate *some* of the problems of needing to reassess decentralized continual self-learning AI systems. It would do so by requiring producers to actively and passively monitor decentralized iterations of their continual self-learning AI system for reported safety concerns, and such reporting could be used to trigger a broad resubmission of those systems for conformity assessment.

[12] See, for example:

* European Commission, "White Paper: On Artificial Intelligence - A European approach to excellence and trust," EN, 19.2.2020, p. 14.

* European Commission, "Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics," EN, 19.2.2020, pp. 10, 14.

the competitiveness of European companies,"[13] so we urge the Commission to reduce this legal uncertainty by formally outlining how stand-alone software is covered in updated directives. In particular, the Expert Group on Liability and New Technologies has identified for the Commission that the "for the purposes of the [Product Liability Directive], products are defined as movable objects, even when incorporated into another movable or immovable object, and include electricity…" and that " emerging digital technologies, especially AI systems, challenge that clear distinction."[14] We recommend that to resolve this uncertainty, the Commission closely consider the recommendations made by the Expert Group and others. For example, it has been suggested that the most effective approach may be to include AI systems as a product under relevant Directives even though it is not a 'movable' object.[15] In the end, from a lay perspective, the manner in which the Commission resolves this uncertainty should *not* alleviate the responsibilities of producers to provide for the safety of their AI systems, and remain liable for damages produced by such systems, whether the AI system is legally defined as a service, product, or else how.

Recommendation: Restrict use of the development risk defence and later defect defence for liability protection on continual self-learning AI systems

As the Commission does, FLI believes firmly in the power of civil liability rules to "play a double role in our society: on the one hand, they ensure that victims of a damage caused by others get compensation and, on the other hand, they provide economic incentives for the liable party to avoid causing such damage."[16] Thus, the Commission must address the unique challenges AI systems provoke for existing civil liability rules to maximize these two benefits for the EU. To that end, FLI agrees with the Expert Group on Liability and New Technologies and their concern about the inappropriateness of the development risk defence for emerging digital technologies, especially AI systems.[17] The development risk defence "allows the producer to avoid liability if the state of scientific and technical knowledge at the time when he put the product into circulation was not such as to enable the existence of the defect to be discovered." [18] Especially in the case of continual self-learning AI, we believe that the state of scientific and technical knowledge is such that it is understood that these AI systems can develop in unforeseen ways if not closely monitored. Thus, it is entirely predictable with current knowledge that unforeseen developments might occur, and therefore the development risk defence does

[13] European Commission, "White Paper: On Artificial Intelligence - A European approach to excellence and trust," EN, 19.2.2020, p. 12.
[14] Expert Group on Liability and New Technologies – New Technologies Formation, "Liability for Artificial Intelligence and other emerging technologies," 21.11.2019, p. 28.
[15] See Wagner, G, "Produkthaftung für autonome Systeme" in Archiv für die civilistische Praxis, p. 718, available at https://www.rewi.hu-berlin.de/de/lf/oe/rdt/pub/working-paper-no-3
[16] European Commission, "White Paper: On Artificial Intelligence - A European approach to excellence and trust," EN, 19.2.2020, p. 12.
[17] Expert Group on Liability and New Technologies – New Technologies Formation, "Liability for Artificial Intelligence and other emerging technologies," 21.11.2019
[18] Ibid, p. 28.

not apply, as identified by the Expert Group.[19] Different member states of the EU already do not use this clause, which demonstrates the feasibility of limiting its application.[20] Furthermore, in contrast to many other goods, a producer can more cheaply address potential defects in AI systems as the producer becomes aware of them through software updates that are less costly than traditional recalls.

There is also a related liability defense if a defect is found to be a 'later' defect (a defect that did not yet exist when the product was given to the customer)**.** As with development risk defence, the later defect defence poses a challenge to an effective regulatory framework for AI systems, and decentralized and centralized continual self-learning AI systems in particular. FLI believes that the 'defects' of continual self-learning AI systems should *not* be classified as 'later' defects if the defect is tied directly to the fact that the learning done by the system in the product lifecycle produced unanticipated harm. The fact that a system can self-learn while being used by a customer is not a 'later' defect, as self-learning is at the very core of the market value of the product that is known to have possible unintended negative consequences.[21]

Despite our recommendation in the above, we acknowledge that relatively few circumstances may arise with continual self-learning AI systems, when a producer should *not* be held liable for future defects in a continual self-learning system. For example, it is possible that a producer has provided a centralized continual self-learning AI system to a customer, who elects to negligently disconnect the system from further updates, including critical safety updates. Or, in the case of a decentralized self-learning AI system, the customer does not allow the producer to provide a periodic safety audit of the system, or intentionally manipulates the AI system through adversarial training examples to produce harmful outcomes. In such circumstances, we believe the EU judicial system can properly adjudicate such circumstances on a case-by-case basis, as it may violate an implied or explicit warranty for the AI system and, therefore, may exempt the producer from liability. These potential cases, which effectively result from a customer's negligent behaviour, should not be used by producers as a means of entitlement to abusing the development risk defence or later defect defence to avoid other liability.

Recommendation: Create an obligation to monitor for continual self-learning AI systems

FLI believes that one of the advantages of effective liability regimes is that they can incentivize or require producers to passively and actively monitor the behaviour of the product. It is especially important to strengthen this legal concept to require producers to responsibly oversee and monitor how continual self-learning AI systems are 'evolving.' Active monitoring could

[19] Ibid, p. 43.
[20] Study for the European Commission: "Analysis of the Economic Impact of the Development Risk Clause as provided by Directive 85/374/EEC on Liability for Defective Products" Contract No. ETD/2002/B5, 2014. Countries like Finland and Luxemburg do not have a development risk clause.
[21] Furthermore, having decentralized and centralized continual self-learning AI systems resubmitted for conformity assessments, as FLI proposes in a prior recommendation, also reduces the ability of a producer to use a 'later' defect clause because the product becomes a 'new' product again once reevaluated.

include requirements such that the producer has to proactively search for potential problems in its AI systems that could lead to harm. Active monitoring could include, for example, a formal system for monitoring product performance in the market, frequent testing of existing products, and the review of state-of-the-art academic publications for new AI safety research that may reveal safety flaws in existing products. This active monitoring could be complemented by passive product monitoring. Passive product monitoring could include, for example, providing customers a platform such as a service hotline to report malfunctions, or developing "bug bounty" programs to incentivize independent auditing for safety. For AI systems and other or digital goods, one possible technical solution is to integrate product monitoring into the system such that an additional part of the software/algorithm monitors the behavior and reports atypical patterns in real time.[22] Both of the concepts of active and passive monitoring are established under German law as the 'product monitoring obligation' (Produktbeobachtungspflicht)[23], and should be carefully evaluated by the Commission for replication. Further, there is another precedent that the Commission should reference for this style of a monitoring framework for products that can present unpredicted risk even after carefully reviewed *prior* to being put into circulation. That precedent is for medicines under the concept of pharmacovigilance, as managed by European Medicines Agency (EMA). Just as the EMA, through the EudraVigilance, works with pharmaceutical manufacturers to monitor for suspected adverse reactions to medicines in the EU, so too should the EU consider developing a mechanism to work with AI system producers to monitor for unanticipated harms.

B. Recommendations on the Risk-Based Approach

FLI agrees with the Commission that a key element of building an 'ecosystem of trust' will come from supporting "rules protecting fundamental rights and consumers' rights, in particular for AI systems operated in the EU that pose a high risk."[24] Thus, a proportionate regulatory framework ought to prioritize AI harms (both material and immaterial) that pose the greatest risks, particularly to safety and fundamental rights. However, AI risks are both broad and mutable, making firm determinations at the outset a challenge. The Commission's current description of the criteria for a high-risk application rightly emphasizes the need to understand the context by looking both at the sector and at the specific use. Nonetheless, a reckless or unsafe AI system within any sector is capable of jeopardizing safety and fundamental rights. We recommend primarily focusing attention on the specific use, and employing a relatively generous scope of potentially risky sectors. Moreover, the inclusion of "exceptional instances, where, due to the

---

[22] For example, this was proposed in Schmid, Alexander, "Pflicht zur „integrierten Produktbeobachtung" für automatisierte und vernetzte Systeme," 19.3.2019, available at https://www.degruyter.com/view/journals/cr/35/3/article-p141.xml

[23] For a description of the Produktbeobachtungspflicht in English, see Günther, J. and Eck, D., "German civil liability for users and manufactures of robotic transportation systems," 11.6.2012, 2012 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), available at https://ieeexplore.ieee.org/document/6213400

[24] European Commission, "White Paper: On Artificial Intelligence - A European approach to excellence and trust," EN, 19.2.2020, p. 3.

risks at stake, the use of AI applications for certain purposes is to be considered as high-risk as such"[25] is an important caveat to the cumulative criteria scheme, and we strongly encourage the Commission to retain this approach in future regulatory frameworks.

Though FLI is generally very appreciative of the risk-based approach the Commission has set forth in the White Paper, we do have several recommendations for improvement.

Recommendation: Include consideration of societal harms in risk assessments

We agree that relevant threats from AI applications include that they can "produce legal or similarly significant effects for the rights of an individual or a company; that pose risk of injury, death or significant material or immaterial damage; that produce effects that cannot reasonably be avoided by individuals or legal entities."[26] We commend the Commission for accounting for this broad range of harms when identifying and evaluating systems that may pose high risk to EU citizens. However, in addition to material and immaterial harms that can occur for individuals, the Commission should also evaluate whether AI applications can cause societal-level harms, even while only producing negligible harms to individuals. For example, AI applications that moderate the information an individual receives through media platforms can meaningfully influence societal consumer choice or democractic decision-making once they reach a sufficiently large number of users. The effect on an individual user may be marginal and impossible to measure (e.g., buying one extra superfluous piece of clothing a year, or reducing an individual's desire to vote by a minor amount), but the effect at a societal level through the aggregation of the effect can be extreme. In particular, applications that could produce widespread disinformation or manipulation of group behaviour, amongst other sociological or cultural effects, should be considered high-risk.

We also call attention to the fact that AI risks do not only emerge from AI systems doing something other than what they are intended to do. In some cases, it can be precisely because an AI system is so effective at carrying out its goals that harm is caused. This can be the case either if the system is designed to achieve nefarious ends, or if it is designed to achieve beneficial ends, but does so in a way that is inadvertently harmful. This risk of AI systems has been recognized by other governments, including the United States, which has stipulated that "there is a risk that AI's pursuit of its defined goals may diverge from the underlying or original human intent and cause unintended consequences — including those that negatively impact privacy, civil rights, civil liberties, confidentiality, security, and safety."[27] Thus, in many regards,

---

[25] Ibid, p. 18

[26] European Commission, "White Paper: On Artificial Intelligence - A European approach to excellence and trust," EN, 19.2.2020, p. 17.

[27] U.S. Office of Management and Budget, "Guidance for Regulation of Artificial Intelligence Applications" (DRAFT), p. 12, available at https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf**.** In coordination with other partner organizations, FLI submitted comments on this regulatory guidance from the U.S. government, available at https://www.regulations.gov/document?D=OMB-2020-0003-0081.

the Commission must also monitor for the possible harms produced by an AI system if it achieves its defined goals *too* well (e.g., such as click maximization).

Recommendation: Create a multi-tiered risk assessment framework

As noted by the Commission, a determination of what constitutes a "high-risk AI application" should be "clear and easily understandable and applicable for all parties."[28] We are concerned that having a strictly binary system of 'high-risk' and no-high risk will strain the ability to appropriately make such clear determinations (even if the criteria are made more explicit. This will potentially result in the vast majority of AI systems remaining under-scrutinized by the EU. Given the polarity of the stakes between 'high' and 'no-high' risk categorizations, companies developing AI applications will have a strong incentive to first minimize what the high-risk category will include, and then to argue that their applications do not fit the criteria. The Commission will no doubt receive numerous comments from companies attempting to achieve this outcome. Creating an AI regulatory framework for the entire EU whereby the vast majority of AI applications are not subject to mandatory requirements is likely to be a missed opportunity, and potentially a dangerous one.

Importantly, a binary risk assessment is not the only option for categorizing AI risks. As the Commission references, the German Data Ethics Commission has called for a five-level risk-based system of regulation based on the criticality of the algorithms.[29] Level one (the base of the pyramid and the majority of use cases) represents AI applications with zero or negligible potential for harm, and these do not require special measures above and beyond existing law. Level five (the tip of the pyramid and the smallest number of use cases) represents applications with an untenable potential for harm. The scheme calls for the complete or partial ban of such algorithmic systems. The three intermediary levels would be regulated to varying degrees. Level two applications, with some potential for harm, would include requirements such as the publication of a risk assessment and monitoring procedures. Level three applications, with regular or significant potential for harm, would require additional measures such as ex-ante approval procedures. Lastly, level four applications, with serious potential for harm, would demand additional measures such as continuous oversight by supervisory institutions.

Obviously, the nuance afforded by a five-tiered risk assessment scheme is significantly greater than a binary risk assessment scheme. It is noteworthy that such a system allows for the recognition that may exist a very small subset of applications that can be so dangerous they should not be allowed in any circumstance. As AI systems advance, we could face a reality in which a lack of human control and predictable AI alignment with goals, for example, demands such measures from the EU. Thus, it would be wise to consider the inclusion of greater nuance to allow for such distinctions. More immediately, having three tiers of AI regulation is likely to

[28] Ibid., p. 17

[29] See Figure 8, on the "Criticality pyramid and risk-adapted regulatory system for the use of algorithmic systems" in Daten Ethik Kommission, "Opinion of the Data Ethics Commission," EN, 13.10.2019, p. 171, available at https://datenethikkommission.de/wp-content/uploads/DEK_Gutachten_engl_bf_200121.pdf

enable the application of more precise and targeted regulatory interventions. For example, ex-ante approval would only be required for applications rising to at least a level three, but the public would still be afforded some oversight over level two applications.

In addition to the German Data Ethics Commission's proposal, another relevant precedent for the Commission to consider is the model of risk classification used in the EU's Medical Device Directive (MDD) and Medical Device Regulations (MDR).[30] This scheme includes four tiers: Class I (low risk), Class IIa (medium risk), Class IIb (medium/high risk), and Class III (high risk). The MDR includes 22 classification criteria to help distinguish between the risk categories, each of which corresponds to varying regulations, controls, and approval pathways for a medical device to be cleared for use in the market. This is similar to the three-tiered risk-stratified approach used in the United States for the regulation of medical devices, including: Class I: Low Risk (with general controls,) Class II: Medium Risk (with both general and special controls,) and Class III: High Risk (with general controls and a premarket approval application. These more nuanced risk management frameworks can better account for the variability of risks, both material and immaterial, that AI systems may pose, which, in general, far exceeds that of medical devices. Thus, the regulatory approach to AI systems would seemingly require as much or greater nuance than medical devices, particularly when you account for the continual self-learning AI systems discussed in the prior section.

Recommendation: Evaluate supplemental governance methods to accompany a voluntary labelling scheme

As currently described by the Commission, AI applications that do not qualify as high-risk will not be subject to any mandatory requirements beyond existing EU rules. The Commission proposes that a voluntary labeling scheme may be able to compensate for this lack of additional oversight. If properly developed and implemented, especially with the "combination of ex ante and ex post enforcement" described by the Commision,[31] FLI supports a voluntary labelling scheme as *one* potential tool among many. The Commission suggests that such a labelling system will "allow users to easily recognize that the products and services in question are in compliance with certain objective and standardised EU-wide benchmarks." While we agree that such a labelling system is possible, the Commission will need to carefully consider how to communicate certain values (e.g., the trustworthiness of an AI system) in a labelling scheme so that they are "easily recognize[d]." We acknowledge that past EU labelling schemes have not always provided easily recognized information to consumers.[32] If consumers do not understand or appreciate the values being communicated by the labels, then the system will provide little in the way of quality assurance for no-high risk products. One way to support both greater

[30] Council Directive 93/42/EEC and Regulation (EU) 2017/745, respectively.
[31] European Commission, "White Paper: On Artificial Intelligence - A European approach to excellence and trust," EN, 19.2.2020, p. 24.
[32] See Ipsos and London Economics, "Consumer market study on the functioning of voluntary food labelling schemes for consumers in the European Union," 12.2013, available at https://ec.europa.eu/info/publications/voluntary-food-labelling-schemes-study_en

transparency by AI developers and improved understanding by consumers is to require certain plain language documentation standards, reviewed on an ongoing basis, as a condition of inclusion in the program.

More fundamentally, a voluntary labelling scheme will only work if there is sufficient, voluntary buy-in from AI technology companies. However, FLI anticipates that the Commission may find that some companies do not wish to participate in such a scheme.[33] Therefore, the Commission may be required to develop alternatives to a voluntary labelling scheme to provide supplementary means of oversight of no-high risk AI systems. For example, the Commission may choose to issue sector-specific, or use-specific, guidance or frameworks with regard to AI systems that are not classified as high risk. Such commentary, while not having the force of law, would still help consumers and consumer advocates evaluate the safety of AI systems not subjected to more rigorous oversight. Another possible option is for the Commission to provide incentives to AI developers to publish risk assessments and invest in monitoring practices. However, any such voluntary measures will be more effective for lower-risk AI applications, while medium-risk AI applications would likely benefit most from the adoption of new regulatory mechanisms, as discussed in the prior recommendation.

C. Recommendations on Fiduciary Responsibilities, Disclosure of Conflicts of Interest, and Loyalty

Our final set of recommendations are thematically unified, and regard concepts largely absent from consideration in the Commission's White Paper or supporting documents. First, we wish to acknowledge that AI systems can provide immense economic and social benefit by, in effect, replacing or substituting for activities previously performed only by a human. For example, expanding the use of AI systems into healthcare settings might lower the cost of certain medical services and help proliferate their use in underserved populations by automating healthcare functions. These functions were previously only performed by highly skilled healthcare workers, and thus had limited supply. At FLI, we do not wish to discourage this application of AI systems. However, these applications entail new challenges for governance, particularly when certain legal restrictions are *assumed* to apply to the AI systems replicating the function of what a human being might do, or previously did.

Thus, FLI recommends the following to address these issues.

---

[33] For example, Google has already publicly indicated in their submitted response to the Commission's White Paper that they would "rather not" have a voluntary labelling system, and that such a system would place a "significant burden on [small and medium enterprises] to comply. This would favour large players who can afford to meet the requirements whilst delivering minimal benefit to consumers." Google's submission is publicly available at https://www.blog.google/documents/77/Googles_submission_to_EC_AI_consultation_1.pdf

Recommendation: Establish reciprocal legal responsibilities for AI systems

An AI system[34] can perform a portion of, or replace entirely, the function of a human actor who has certain legal responsibilities to another individual. In such circumstances, the legal responsibilities held by the substituted human actor should be reciprocated by the responsible agent making use of the AI system. For example, if a financial company begins using an AI system to recommend a financial course of action to a customer, the financial company should have the same legal responsibilities to the customer as if the company would have used a human employee to perform this task. Thus, in this example, the financial company's AI system should comply with the 'prudent person principle' or related fiduciary responsibility in a reciprocal manner that is called for under relevant EU or Member State law. Likewise, AI systems providing or recommending a course of treatment to a patient in a manner similar to a human doctor should comply with numerous legal variations of patient privacy. We acknowledge that implementing such a sweeping set of reciprocal responsibilities could be difficult to harmonize at the EU level. Therefore, this idea could be put forth as a broadly worded directive as opposed to a regulation, to allow for Member States to implement it in context specific ways. However, it can readily be assumed that failure to do so will result in numerous instances where European consumers will *assume*, incorrectly, that legal responsibilities they expect their doctor, lawyer, or other provider to have toward them as consumers will convey to the use of products and services enabled by AI that perform nearly identical functions. This will violate not only the trust of the European consumers, but result in potentially grave immaterial, and potentially material, harms.

Recommendation: Mandate disclosures of conflicts of interests in AI systems

There are many AI systems, especially digital services, where consumers reasonably expect that the information provided by the AI system meets certain unstated standards. For example, if someone uses a GPS and AI-enabled mobile mapping service to develop driving directions from Point A to Point B, it may be reasonable to expect the directions provided are the shortest, or fastest, route. This expectation is largely the same as if someone were to ask a friend, or a stranger, for directions between Point A and Point B (the person asked would likely try to provide the best possible directions to you, unless they are acting maliciously). However, this expectation can be violated covertly by an AI-enabled mapping service, as it might also include subtle weights in the algorithm that increase the chances that the directions will lead you past a certain type of business. This behaviour by the AI system subsequently may increase the number of individuals that stop to purchase a meal or other service from the business, unbeknownst to the users of the mapping service. It may be reasonable, especially for free services, that the producer of the AI system needs to have some ability to gain monetary benefit from the system through such manipulations. However, these conflicts of interest should be disclosed to the consumer.

[34] This should be done regardless of whether the AI system is classified as a product or service under EU law.

For the above and other similar circumstances, there should be mandatory disclosures about inherent conflicts of interest in AI products/services that do not have a reciprocal legal responsibility already placed upon their use. For example, the Commission could require producers to label AI systems as having a conflict of interest, or bias, to produce various outcomes that do not serve the customer's interest alone. Producers of such systems will undoubtedly be aware of such conflicts, as they are likely driven by a market or financial incentive, and should be able and required to disclose their existence.

Recommendation: Evaluate AI "loyalty" as an immaterial risk in conformity assessments

As previously mentioned, FLI fully supports the Commission's intent to include both material and immaterial harms from AI systems in a future regulatory approach. The Commission identifies examples of immaterial harms such as the "loss of privacy, limitations to the right of freedom of expression, human dignity, discrimination, for instance in access to employment."[35] We acknowledge that the Commission did not intend to provide an exhaustive accounting of all possible immaterial harms in this context. In a future regulatory approach, however, FLI believes it is important to specifically consider a more novel, and less recognized, set of immaterial harms that can arise from an AI system that is "disloyal" to its primary owner/customer. An AI system is loyal "to the degree that they are designed to minimize, and make transparent, conflicts of interest, and to act in ways that prioritize the interests of users."[36] The prior recommendations of this section reinforce portions of what it means for an AI to be 'loyal' to a user or customer. However, it may not cover all possible harms for an AI system that is 'disloyal.'

Therefore, immaterial harm from disloyal behavior by AI systems should be considered in conformity assessments of high-risk AI systems. In considering this factor, regulators should be especially mindful of the potential problems and risks that may arise when an AI system is *assumed* to be loyal with a particular set of utility functions, but is actually designed to maximize utility functions unknown to the user (or perhaps even the regulator) in a hidden, disloyal way. For instance, a personalized educational AI system may suggest to a user a particular curriculum or set of academic articles to review on a topic, with the *assumed* function of providing resources the user would most benefit the user's learning. Rather, the education system is actually trying to persuade the user to adopt a particular set of opinions about a relevant topic, such as the health effects of smokeless nicotine products or the economic effects of lower taxation rates, to the benefit of paying sponsors for the producer of the AI system. In doing so, the AI system may create immaterial harm, both by deceiving the user of the educational system and by increasingly skewing their thinking toward a particular understanding

[35] European Commission, "White Paper: On Artificial Intelligence - A European approach to excellence and trust," EN, 19.2.2020, p. 10.
[36] Aguirre, A., Dempsey, G., Surden, H., and Reiner, P.B., "AI loyalty: A New Paradigm for Aligning Stakeholder Interests," 25.3.2020, U of Colorado Law Legal Studies Research Paper No. 20-18. Available at SSRN: https://ssrn.com/abstract=3560653

on a topic. This disloyalty can cause further immaterial harm, such as undermining the functioning of society dependent on a well-educated population with a diversity of viewpoints.